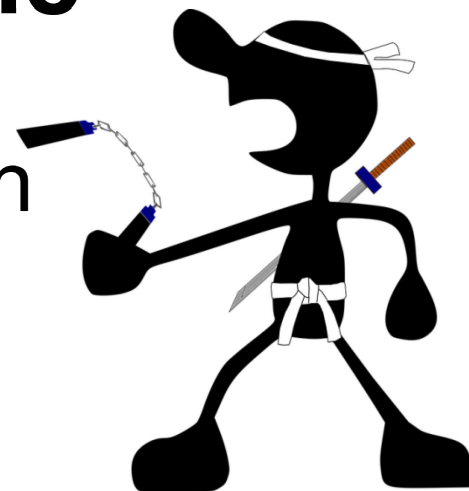
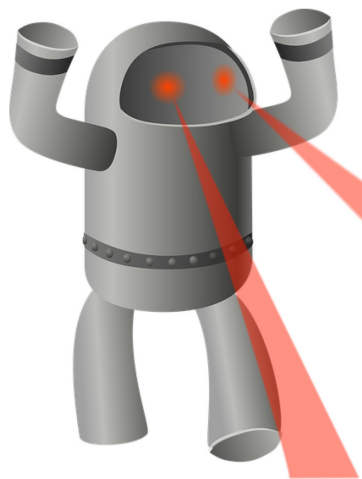


Mensch vs. Maschine

Texterfassungsmethoden
auf dem Prüfstand



13.06.2018



Dr. Irene Schumm
Dr. Philipp Zumstein

Mensch vs. Maschine – Agenda

- Text- und Strukturierung in Druckschriften
- DFG-Projekt Aktienführer
- Vergleich der Text- und Strukturierungsmethoden anhand ausgewählter Kriterien

Digitalisierung von Druckschriften

Scannen	✓
Erschließung der bibl. Metadaten und des Inhaltsverzeichnisses	✓
Volltexterfassung	(✓)
Erfassung von Strukturen im Volltext	(✓)
Präsentation	✓

Beispiel Digitalisierung Aktienführer

- Nachschlagewerk für Informationen zu Aktiengesellschaften
- DFG-Projekte zur Digitalisierung der Jahrgänge 1870-2016
- Strukturierte Volltexterfassung für die Jahrgänge 1953-1999
- Vorlagenspezifische Gegebenheiten (Struktur, Zahlen, Tabellen)

[illegible]

BASF Aktiengesellschaft

BASF Sitz: 6700 Ludwigshafen (Rhein)
Telefon: (06 21) 8 01
Telex: 4 04 811

Dr. jur. Robert Ehret, Eltingstein/Trause;
Prof. Dr. rer. nat. Manfred Elgen,
Göttingen;
Prof. Dr.-Ing. Berthold Frank, Heidelberg;
Dr. rer. pol. Johan M. Grootenwaard,

Yarabandi:

[illegible]

Veränderungen auf GuV-Rechnungen			1976		
	1976	1977		1976	1977
1976: 8 % (Bw. Sch. Nr. 24)			Fremdkapital	3 300	10 200
1974 + 1975: je 4 % (Bw. Sch. Nr. 25, 26)			Umsatzgewinn	240	240
1976: 8 % (Bw. Sch. Nr. 27)			Bilanzsumme	14 655	16 090
1977: 6 % - 1978: 3, 20 % (Bw. Sch. Nr. 28)					

Aus den Bilanzen (in 1000 DM)			Aus den Gewinn- und Verlustrechnungen		
	1976	1977		1976	1977
Umsatzerlöse	4 395	5 300	Umsatzerlöse	17 810	22 120
Umsatzkosten	1 910	10 240	Materialeinzelkosten	2 510	8 700
Umsatzgewinn	2 485	12 240	Umsatzkosten	15 300	13 420
Umsatzkosten	2 410	11 240	Umsatzkosten	810	740
Umsatzkosten	2 410	11 240	Umsatzkosten	810	740
Umsatzkosten	2 410	11 240	Umsatzkosten	810	740

BASF Aktiengesellschaft

BASF SITE: 6900 Ludwigshafen (Rhein)
Telefon: (06 21) 8 01
Telefax: 4 64 811

Dr. rer. Robert Ehret, Kohnstein/Taunus
Prof. Dr. rer. nat. Manfred Eigen, Göttingen
Prof. Dr.-Ing. Berthold Frank, Heidelberg
Dr. rer. pol. Johan M. Goudswaard,

Verstärkt:

[illegible]

Aus den Bilanzen		
	31.12.1953	31.12.1954
	(in 1 000 DM)	
Aktiva		
Langfristige Aus-		
leihungen	96 919	161 048
Ausgleichsfor-		
derungen	4 504	8 869
Wertpapiere	9 051	15 788
Barmittel	27 297	28 349

Vorstand:

Prof. Dr. jur. Reimer Schmidt, Aachen,
Vors.;

Dr. jur. Helmut Gies, Aachen;

Dr. rer. nat. Christian Hammer, Stolberg;

Dr. jur. Bernd Michaels, Aachen;

Carl Putens, Aachen;

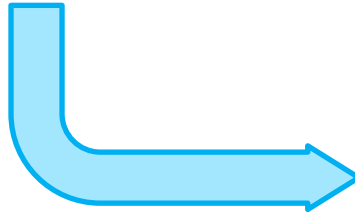
Dr. jur. Dr. rer. pol. Johannes Schießl,
Aachen;

- <https://digi.bib.uni-mannheim.de/aktienfuehrer/>

Motivation Texterfassung

- Recherchemöglichkeiten in den Digitalisaten
- Kopiermöglichkeiten für den Volltext

Aktiengesellschaft




			B	
Dividenden auf Stammaktien:			1976	1977
1973: 8 % (Div. Sch. Nr. 24)			Fremdkapital	9 360 10 833
1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26)			Bilanzgewinn	242 185
1976: 8 % (Div. Sch. Nr. 27)			Bilanzsumme	14 655 16 094
1977: 8, - DM + 3,38 DM St. G. (Div. Sch. Nr. 28)			Aus den Gewinn- und Verlust- rechnungen	
Aus den Bilanzen (in 1 000 DM)			1976	1977
			1976	1977
Anlagevermögen	4 395	5 305	Umsatzerlöse	17 910 22 106
Umlaufvermögen	10 260	10 788	Materialaufwand	7 510 8 579
(flüssige Mittel)	519	727	Personalaufwand	9 237 10 389
Eigenkapital	5 117	5 117	Abschreibungen	810 767
(Grundkapital)	3 015	3 015	EEV-Steuern	299 539
			Jahresüberschuß	241 184

BASF Aktiengesellschaft	
	Sitz: 6700 Ludwigshafen (Rhein)
	Telefon: (06 21) 6 01
	Telex: 4 64 811
Vorstand:	
Prof. Dr. rer. nat. Matthias Seefelder, Ludwigshafen (Rhein), Vors.:	
Dr. jur. Robert Ehret, Königstein/Taunus;	
Prof. Dr. rer. nat. Manfred Eigen, Göttingen;	
Prof. Dr.-Ing. Berthold Frank, Heidelberg;	
Dr. rer. pol. Johan M. Goudswaard, Wassenaar (Niederlande);	
Dr. jur. Wolfgang Heintzeler, Heidelberg;	
Kurt Herrmann, Carlsberg/Pfalz *);	
Dr. rer. pol. Kurt Hohenemser, Frankfurt;	
Dr. jur. Robert Holzach, Zumikon (Schweiz);	

Motivation Strukturerfassung

- Passgenaue Daten zur Beantwortung von Forschungsfragen
- Ohne viel blättern oder selbst Daten erfassen zu müssen



Frauenquote in den
Vorständen der DAX-
30-Unternehmen in
den letzten
Jahrzehnten?

Vorstand:

Prof.Dr.jur. **Reimer** Schmidt, Aachen,
Vors.;
Dr.jur. **Helmut** Gies, Aachen;
Dr.rer.nat. **Christian** Hammer, Stolberg;
Dr.jur. **Bernd** Michaels, Aachen;
Carl Putens. Aachen;



Name	Vorname	Titel	Ort	Funktion
Schmidt	Reimer	Prof. Dr. jur.	Aachen	Vors.
Gies	Helmut	Dr. jur.	Aachen	
Hammer	Christian	Dr. rer. nat.	Stolberg	

Text- und Strukturierungsmethoden im Überblick

Manuell



Aktienführer I

Jahrgänge:
1976-1999

Maschinell



Aktienführer II

Jahrgänge:
1953-1975

Vergleich der Text- und Strukturermassungsmethoden

Im Folgenden werden die beiden Methoden anhand folgender Kriterien verglichen:

- Ressourcen
- Organisation
- Nachnutzbarkeit
- IT-Kenntnisse und -Tätigkeiten im Haus
- Güte Text- und Strukturerkennung
- Typische Erkennungsfehler

Ressourcen

Manuell

- 16 Personenmonate
- Organisat. Know-how
- Studentische Hilfskräfte
- Kosten für das Double Keying

Maschinell

- 2 x 12 Personenmonate
- IT Know-how
- Studentische Hilfskräfte
- Rechnerinfrastruktur und Lizenzen

Organisation

Manuell

- Erfassung außerhalb
- Pflichtenheft erstellen für Dienstleister
- Vorbereitung & Durchführung von Ausschreibung, Auftragsvergabe
- Begleitung des Erfassungsprozesses (Rückfragen durch den Dienstleister)
- Kontrolle, Reklamation, Nachbesserung und Auftragsabnahme

Maschinell

- Erfassung im Haus
- Internes IT-Projektmanagement
 - Planung und Steuerung der Softwareentwicklung
 - Bereitstellung der Software als Open Source

Nachnutzbarkeit

Manuell

- Digitalisate und Daten

Maschinell

- Digitalisate und Daten
- Weiterentwicklung OCR-Engines
- Neue Open Source Tools
 - [CRASS](#) (crop & splice segments)
 - [ocromore](#) (Multi-OCR-Verfahren)
 - mocrin (Koordination OCRs)
 - Strukturparser
 - ...



IT-Kenntnisse und -Tätigkeiten im Haus

Manuell

- ∅
- Datenbankerstellung und Online-Präsentation

Maschinell

- Einarbeitung in verschiedene OCR-Engines
- Software-Entwicklung
- Algorithmen-Entwicklung
- Bildbearbeitung, Layout-Analyse, Texterkennung, Parsing
- Datenbankerstellung und Online-Präsentation

Güte Text- und Strukturerkennung

Manuell

- Vorgabe an den Dienstleister: 99,90 % Erfassungsgenauigkeit
- Stichprobenartige Überprüfung der Ergebnisse nach Text- und Strukturerrfassung
- Vergessene Profile und Datenkategorien
- „maschinenunlesbare“ Daten

Maschinell

- Stichprobenartige Überprüfung der Ergebnisse nach Texterkennung
- Vor Triple-OCR: zwischen 84,1 % und 98,8 %
- Nach Triple-OCR: 99,2 %
- Bilder in Übergrößen überfordern teilweise OCR-Engines

Typische Erkennungsfehler

Manuell

- Währung, z.B. \$ £ ₪ ¢ fl
- Leerzeichen bei Zahlen

6 1/2 % -> 61/2%

- Invalides XML-Konstrukt

<Betrag>129</Betsrag>

- Strukturierungsfehler

```
Aufsichtsrat15>
  <Name>von Thurn und Taxis</Name>
  <Vorname>Johannes</Vorname>
  <Titel>Erbprinz</Titel>
</Aufsichtsrat15>
```

Maschinell

- Interpunktation (~50 %), z. B.

Prof. Dr. rer. nat.

Prof. Dr. rer, nat.

- Akzente, Umlaute, z. B.

BASF Española

BASF Espanola

- Abstandsfehler

- Druckfehler

1 100 000 000.-

Fazit

- Resultate der manuellen und maschinellen Text- und Strukturerkennung sind **qualitativ ähnlich gut**
- Manuelles Verfahren funktioniert auch **ohne umfangreiches IT-Know-how**
- **Ressourcenaufwand** beider Verfahren **ähnlich**
- **Nachnutzbarkeit der Methoden** beim maschinellen Verfahren (Open Source) für nächste Projekte
- Algorithmen und Software werden sich weiterentwickeln, z. B. durch OCR-D-Projekte (**Verbesserungspotentiale beim maschinellen Verfahren**)

Bildquellen

- Folie 1: <https://pixabay.com/de/angriff-todesstrahl-b%C3%B6se-laser-1294254/> (CC0),
<https://pixabay.com/de/k%C3%A4mpfer-martial-arts-asiatische-1293871/> (CC0)
- Folie 5: <https://iconmonstr.com/magnifier-2-png/> (Licensee may use the Work in non-commercial and commercial projects, services or products without attribution.)
- Folie 7: <https://pixabay.com/de/programmierer-programmierung-code-1653351/> (CC0)